A Brief Overview on the Strategies to Fight Back the Spread of False Information

Álvaro Figueira[1], Nuno Guimaraes[1] and Luis Torgo[2]

[1]*CRACS-INESCTEC and University of Porto, Porto, Portugal*
[2]*Faculty of Computer Science, Dalhousie University, Halifax, Nova Scotia, Canada*
E-mail: arf@dcc.fc.up.pt; nuno.r.guimaraes@inesctec.pt; ltorgo@dal.ca

Received 09 January 2019;
Accepted 07 June 2019

Abstract

The proliferation of false information on social networks is one of the hardest challenges in today's society, with implications capable of changing users perception on what is a fact or rumor. Due to its complexity, there has been an overwhelming number of contributions from the research community like the analysis of specific events where rumors are spread, analysis of the propagation of false content on the network, or machine learning algorithms to distinguish what is a fact and what is "fake news". In this paper, we identify and summarize some of the most prevalent works on the different categories studied. Finally, we also discuss the methods applied to deceive users and what are the next main challenges of this area.

Keywords: false information, social networks.

Journal of Web Engineering, Vol. 18_4-6, 319–352.
doi: 10.13052/jwe1540-9589.18463

1 Introduction

The large increase of social media users in the past few years has led to an overwhelming quantity of information available in daily (or even hourly) basis. In addition, the easy accessibility to these platforms whether it's by a computer, tablet or mobile, allows the consumption of information at a distance of a click. Therefore, traditional and independent news media urge to adopt social media to reach a broader audience and gain new clients/consumers.

The ease of creating and disseminating content in social networks like Twitter and Facebook has contributed to the emergence of malicious users. In particular, users that infect the network with the propagation of misinformation or rumors. These actions combined with the fact that 67% of adults consume some type of news in social media (20% on a frequent basis) [24] have already caused real-world consequences [56].

However, unreliable content or, how it is now referred – "fake news" –, is not a recent problem. Although the term gained popularity in the 2016 US presidential election, throughout the years newspapers and televisions have shared false content resulting in severe consequences for the real world. For example, in 1924 a forged document known as "The Zinoviev Letter" was published on a well known British newspaper four days before the general elections. The goal was to destabilize the elections in favour of the conservative party with a directive from Moscow to British communists referring an Anglo-Soviet treaty and inspiring "agitation-propaganda" in the armed forces [39]. Another example happened after the "Hillsborough accident", where 96 people died crushed due to overcrowding and lack of security. Reports from an illustrious newspaper claimed that, as people were dying, some fellow drunk supporters stole from them and beat police officers that were trying to help. Later, however, such claims were proven false [15].

The verified impact of fake news in society throughout the years and the influence that social networks currently have today forced high reputation companies, such as Google and Facebook, to start working on a method to mitigate the problem [28, 29]. The scientific community has also been increasing the activity on the topic. In fact, if we search

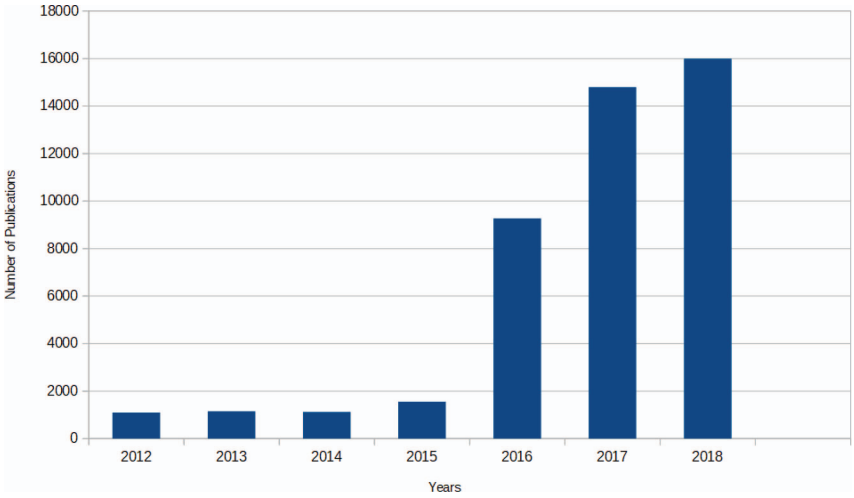

Figure 1 Number of hits per year in Google Scholar for the term "fake news".

in Google Scholar[1] for "fake news", we will find a significantly high number of results that have an increase of 7K publications, when compared with the number obtained in the previous year. Figure 1 illustrates the growth of publication regarding the term "fake news" over the years.

Nevertheless, the problem of fake news is still a reality since the solution is anything but trivial. Moreover, research on the detection of such content, in the context of social networks, is still recent. Therefore, in this work we attempt to summarize the different and most promising branches of the problem as well as the preliminary proposed solutions in the current literature.

In addition, we present a perspective on the next steps in the research with a focus on the need to evaluate the current systems/proposals in a real-world environment.

In the next section, we will present a review of the state of the art on the problem of unreliable information namely the different approaches that are being studied as well as the different kinds of data that is being used. In Section 3 we discuss the problem of unreliable information. Next, we introduce some feature guidelines for future research on

[1]https://scholar.google.com

unreliable content by pointing some limitations on the current solutions proposed in Section 2. Finally, in Section 5, we present the main conclusions of this work.

2 Literature Review

From many well-known examples in the last years, we know that users share on social media first-hand reports (accounts) of large-dimension on-going events, like natural disasters, public gatherings and debates, active shootings, etc.

However, in some occasions not all of what is reported is real. Twitter started to help their users by determining who "is credible" by adding a verified account indicator, which confirms if that account "of public interest" is authentic [27].

Nevertheless, even though the accounts for most genuine users are not verified, many of their social media posts may still be factual. Identifying which users are not credible is thus an important and on-going challenging problem. For example, 'Bots' are often used to spread spam or to rapidly create connections to other users. Bots make it possible to gain followers or "friends" in order to make the user appear more popular or influential [19]. Moreover, they can also flood social media with fake posts to influence the public opinion.

Misinformation (false information) is information that may or may not have been debunked at the time of sharing while rumors are pieces of information which are unverified at the time of sharing and seem to be relevant to a given situation. According to [34], rumors' purpose is to make sense and manage risk situations when there is ambiguity or a potential threat. Curiously, a very common form of disinformation is the dissemination of a "rumor" disguised as a fact.

Misinformation is also conveyed in 'politics' in the form of "Astro-turf campaigns", a collective action where several different accounts spread and promote a false message during a political campaign. Interestingly, most of the times these actions are planned to carefully shift the public opinion [47].

Another form of creating information threats is by the creation of "sybils" to gain importance or authority in the spread of misinformation.

Sybils are similar accounts, in the sense of having a very similar account name, with the intention to fool users by making them believe the account is from some well-known person, friend or entity. Generally, the fake account tries to connect with the friends of the real user/entity. If it is granted then the Sybil account can take advantage of the real user's reputation to more easily spread disinformation. In most cases, once a Sybil account is created, hackers may utilize a set of bots to publish fake messages on social media and reach a much wider, eventually focussed/specific, audience [50].

Another form of misleading information is 'hashtag hijacking', which occurs when a set of people use common hashtags in a context that substantially differs from the original intent [64]. Notice that on social media, astroturfing and hashtag-hijacking are the main modes of *drifting* the public opinion [31].

There are several approaches to tackle the problem of unreliable content on social media. For example, some authors opt by analyzing the patterns of propagation [35, 49, 60], others by creating supervised systems to classify unreliable content [5, 42, 59], and others by focusing on the characteristics of the accounts that share this type of content [7, 9, 17]. In addition, some works also focus on developing techniques for fact-checking claims [12, 52] or focus on specific case studies [2, 13, 26, 57].

In the following subsections, we present a narrative literature review on the current state of the art. This literature review relies on previous knowledge of the authors on the distinct sub-areas of unreliable information on social media. This knowledge was complemented with searches in two different academic search engines: Google Scholar[2] and Semantic Scholar[3]. The queries used in both engines included the key terms "fake news", "misinformation online", "false information" and were combined with the terms "analysis", "detection" and "propagation". Only computer science papers were considered. Literature reviews were excluded from the process. We relied on the search engines to provide the most relevant papers. However, some selection was made to balance the different sub-areas.

[2]http://scholar.google.com

[3]https://www.semanticscholar.org/

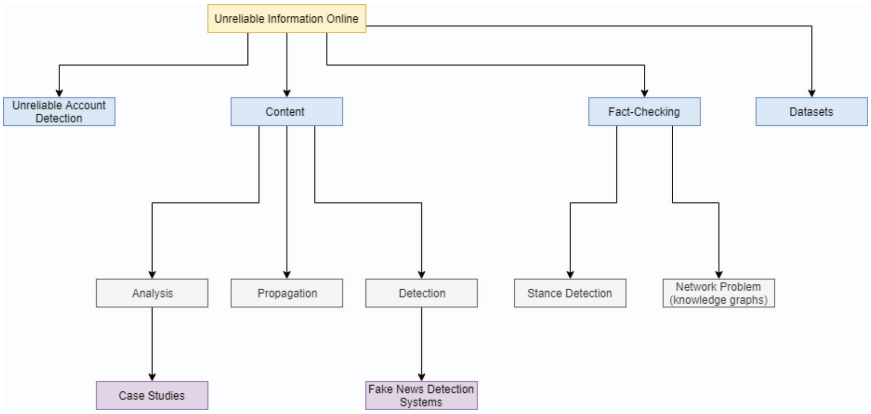

Figure 2 Structure of the state of the art on unreliable content.

This selection was first based on the publishing date (more recent papers were prioritized) and second on the number of citations. In the following subsections, we will provide further detail on the approaches previously mentioned (illustrated hierarchically in 2) and the selected papers for these approaches.

2.1 Unreliable Accounts Detection

Castillo et al. [9] target the detection of credibility in Twitter events. The authors created a dataset of tweets regarding specific trending topics. Then, using a crowd sourcing approach, they annotated the dataset regarding the credibility of each tweet. Finally, they used four different sets of features (Message, User, Topic, and Propagation) on a Decision Tree Model that achieved an accuracy of 86% from a balanced dataset. A more recent work [17] used an Entropy Minimization-Discretization technique that combines numerical features with assessing fake accounts on Twitter.

Benevuto et al. [7] developed a model to detect spammers by building a manual annotated dataset of 1K records of spam and non-spam accounts. Then, they extracted attributes regarding content and user behaviour. The system was capable of detecting correctly 70% of the spam accounts and 96% of non-spam.

A similar problem is the detection of bot accounts. Chu et al. [11] distinguished accounts into three different groups: humans, bots and cyborgs. Using a human-labelled dataset of 6K users, they built a system with four distinct areas of analysis (entropy measures, spam detection, account properties, and decision making). The performance of the system was evaluated using accuracy, which reached 96% in the "Human" class. Another similar work [16], introduced a methodology to differentiate accounts into two classes: humans and bots.

Gilani et al. [22] presented a solution to a similar goal: to distinguish automated accounts from human ones. However, they introduced the notion that "automated" is not necessarily bad. Using a dataset containing a large quantity of user accounts, the authors divided and categorized each entry into 4 different groups regarding the popularity (followers) of the account. The evaluation was conducted using the F1-measure. The results obtained fall between 77% and 91%.

Therefore, regarding the analysis of social media accounts, the majority of the state of the art has been focusing on trying to identify bot or spammer accounts. However not all users that spread unreliable accounts are bots or spammers. Thus, it would be important to analyze those type of accounts also (i.e. human operated accounts that spread unreliable content on social networks). However such studies, to best of our knowledge, are still in a very early stage.

2.2 Content Analysis

Another major area of study is the analysis of large quantities of fake news spread through social networks. Vosoughi et al. [66] presented a study of the differences between propagation of true and false news. The work focused on the retweet propagation of false, true, and mixed news stories for a period of 11 years. The findings were several. First, false news stories peaks were at the end of 2013, 2015 and 2016. Then, through the analysis of retweets of false news stories, the authors concluded that falsehood reaches a significantly larger audience than the truthful. In addition, tweets containing false news stories are spread by users with fewer followers and friends, and that are less active than users who spread true news stories. Another work [65]

studied the agenda-setting relationships between online news media, fake news, and fact checkers. In other words, if each type of content is influenced by the agenda of others. The authors found out that certain issues were transferred to news media due to fake news (more frequently in fake stories about international relations). Furthermore, fake news also predicted the issue agenda of partisan media (more in the liberal side than the conservative). Other relevant findings are the reactive approach of fake news media to traditional and emerging media and the autonomy of fact-checking websites regarding online media agendas.

2.3 Case Studies

Some works focus on analyzing the dissemination of false information regarding a particular event. One of those is related to the Boston Marathon in 2013, where two homemade bombs were detonated near the finish of the race [13]. For example, in [26] the authors performed an analysis on 7.9 million tweets regarding the bombing. The main conclusions were that 20% of the tweets were true facts whether 29% were false information (the remaining were opinions or comments), it was possible to predict the virality of fake content based on the attributes of the users that disseminate it, and accounts created with the sole purpose of disseminating fake content often opt by names similar with official accounts or names that explore the sympathy of people (by using words like "pray" or "victim"). Another work has the analysis focused on the the main rumors spread on Twitter after the bombings occurred [57].

A different event tackled was the US Presidential Election in 2016. For example, the authors in [2] combined online surveys with information extracted from fact-checking websites to perceive the impact of fake news in social media and how it influenced the elections. Findings suggest that articles containing fake news pro-Trump were shared three times more than articles pro-Clinton and the average American adult has seen at least one fake news stories on the month around the election. Another work [8] studied the influence of fake news and well know news outlets on Twitter during the election. The authors collected approximately 171 million tweets in the 5 months prior to the elections

and showed that bots diffusing unreliable news are more active than the ones spreading other types of news (similar to what was found in [2]). In addition, the network diffusing fake and extreme bias news is denser than the network diffusing center and left-leaning news. Other works regarding this event are presented in [33, 51].

Other works address similar events such as Hurricane Sandy [5], the Fukushima Disaster [63] and the Mumbai Blasts in 2011 [25].

2.4 Network Propagation

In Shao [49] the authors expose a method describing the extraction of posts that contained links to fake news and fact-checking web pages. Then, they analyzed the popularity and patterns of the activity of the users that published these type of posts. The authors concluded that users that propagate fake news are much more active on social media than users that refute the claims (by spreading fact-checking links). The authors' findings also suggest that there is a small set of accounts that generate large quantities of fake news in posts.

Another work by Tambuscio et al. [60] describes the relations between fake news believers and fact-checkers. The study resorts to a model commonly used in the analysis of disease spreading, modyfing it and treating misinformation as a virus. Nodes on the network can be believers of fake news, fact-checkers or susceptible (neutral) users. Susceptible nodes can be infected by fake news believers although they can "recover" when confronted with fact-checking nodes. By testing their approach in 3 different networks, the authors concluded that fact-checking can actually cancel a hoax even for users that believe, with a high probability, in the message.

A similar approach is proposed in [35] where a Dynamic Linear Model is developed to timely limit the propagation of misinformation. The model differs from other works since it relies on the ability for the user's susceptibility to change over time and how it affects its dissemination of information. The model categorizes users in 3 groups: infected, protected and inactive, and validates the effectiveness of the approach on a real-world dataset.

2.5 Unreliable Content Detection

A work by Antoniadis et al. [5] tried to identify misinformation on Twitter. The authors annotated a large dataset of tweets and developed a model using the features from the Twitter text, the users, and the social feedback it got (number of retweets, number of favourites, number of replies). Finally, they assessed the capability of the model in detecting misinformation in real time, i.e. in *a priori* way (when the social feedback is not yet available). Evaluations on real-time only decay in 3% when compared with the model that uses all available features. An approach also using social feedback was presented by Tacchini et al. [59]. The authors claim that by analyzing the users who liked a small set of posts containing false and true information, they can obtain a model with an accuracy near 80%.

Perez-Rosas [42] created a crowd-sourced fake news dataset in addition to fake news available online. The dataset was built based on real news. In other words, crowd-source workers were provided with a real news story and were asked to write a similar one, but false. Furthermore, they were asked to simulate journalistic writing. The best model obtained a 78% accuracy in the crowd-sourced dataset and only less 5% in a dataset obtained by fake news on the web.

Another example is the work of Potthast [45] which analyses the writing style of hyper-partisan (extremely biased) news. The authors adopt a meta-learning approach ("unmasking") from the authorship verification problem. The results obtained show models capable of reaching 78% in F1-measure in the task of classifying hyper-partisan and mainstream news, and 81% in distinguishing satire from the hyper-partisan and mainstream news. However, we must note that using only style-based features does not seem to be enough to distinguish fake news since the authors' best result was 46%.

2.6 Fake News Detection Systems

The majority of the implementations to detect fake news comes in the form of a browser add-on. For example, the bs-detector [62] flags content in social media in different categories such as clickbait, bias,

conspiracy theory and junk science. To make this evaluation, the add-on uses OpenSources [40] which is a curated list of dubious websites. A more advanced example is the Fake News Detector [10]. This add-on uses machine learning techniques in a ground truth dataset combined with the "wisdom of the crowd" to be constantly learning and improving the detection of fake news. An interesting system that also took the shape of an add-on was the one developed by four colleges students during a hackathon at Princeton University [4]. Their methodology combined two different approaches: the first makes a real-time analysis of the content in user's feed. The other notifies the user when they are posting or sharing doubtful content. The system is capable of analyzing keywords, recognizes images and verified sources to accurately detect fake content online. With confidence we can say that new systems are being created with a frequency of more than a dozen a year. Most of them uses the add-on approach, but many are not yet suited to be usable by the normal people as they are clearly proof of concept prototypes.

2.7 Fact-Checking

Another way to tackle the problem of false information is through fact-checking. Due to the enormous quantity of information spread through social networks, the necessity to automatize this task has become crucial. Automated fact-checking aims to verify claims automatically through consultations and extraction of data from different sources. Then, based on the strength and stance of reputable sources regarding the claim, a classification is assigned [14]. This methodology, despite being in development is very promising.

2.7.1 Stance Detection

In earlier research, stance detection has been defined as the task of a given fragment of text agrees, disagrees or is unrelated to a specific target topic. However, in the context of fake news detection, stance detection has been adopted as a primary step to detect the veracity of a news piece. Simply putting it, to determine the veracity of a news article, one can look to what well-reputed news organizations are writing about that topic. Therefore, stance detection can be applied to understand

if a news written from an unknown reputation source is agreeing or disagreeing with the majority of the media outlets. A conceptually similar task to stance detection is textual entailment [43, 53]

The Fake News Challenge[4] promotes the identification of fake news through the used of stance detection. More specifically, given a headline and a body of text (not necessarily from different articles), the task consists in identifying if the body of text agrees, disagrees, discusses or its unrelated with the headline. Several approaches were presented using the dataset provided. The authors in [38] present several approaches using a conditioned bidirectional LSTM (Long Short Term Memory) and the baseline model (GradientBoosted Classifier provided by the authors of the challenge) with an additional variation of features. As for the features, Bag of Words and GloVe vectors were used. In addition, global features like binary co-occurrence in words from the headline and the text, polarity words and word grams were used. The best result achieved was using bidirectional LSTM with the inclusion of the global features mentioned. The improvement regarding the baseline was 9.7%. Other works with similar approaches were proposed [43, 53] however, results do not vary significantly.

Stance detection is an important step towards the problem of fake news detection. The fake news challenge seems to be a good starting point to test possible approaches to the problem. Furthermore, the addition of source reputation regarding topics (p.e. politics) can provide useful insight to detect the veracity of a news.

2.7.2 Fact-checking as a Network Problem

The authors in [12] tackle fact-checking as a network problem. By using the Wikipedia infoboxes to extract facts in a structured way, the authors proposed an automatic fact-checking system which relies on the path length and specificity of the terms of the claim in the Wikipedia Knowledge Graph. The evaluation is conducted in statements (both true and false) from the entertainment history and geography domains (for example "x was marry to y", "d directed f" and "c is the capital of r") and an independent corpus with novel statements annotated

[4] http://www.fakenewschallenge.org/

by human raters. The results of the first evaluation showed that true statements have higher truth values than false. In the second evaluation, the values from human annotators and the ones predicted by the system are correlated.

Another work by the same authors [52] use an unsupervised approach to the problem. The Knowledge Stream methodology adapts the Knowledge Network to a flow network since multiple paths may provide more context than a single path and reusing edges and limiting the paths where they can participate may limit the path search space. This technique, when evaluated in multiple datasets, achieves results similar to the state of the art. However, in various cases, it provides additional evidence to support the fact-checking of claims.

2.8 Datasets For Unreliable News Analysis

One key feature to develop reliable systems capable of detecting false information (and the accounts that propagate it) is the quality of the datasets used. In this section, we present some examples of datasets along with their characteristics.

2.8.1 OpenSources

OpenSources [40], as previously mentioned, is a curated dataset of websites that publish different types of unreliable content. Each source is analyzed by experts considering a strict methodology that includes the title/domain analysis, the identification of the owners of the site (through the "about us" section), verification of several articles (by source quotation, or studies to back up the claims made) and verification of writing style, aesthetic and social media presence (i.e. if the posts from the sources' social media page are propitious to clickbait). Each website is characterized to a limit of three different labels. This labels are fake, satire, bias, conspiracy, rumor, state (sources operating under governmental control), junksci (also known as junk science), hate, clickbait, political or reliable.

2.8.2 Fake News Corpus

The "Fake News Corpus" dataset [58] contains approximately 9.5M news articles extracted from the source available in OpenSources

combined with the New York Times and Webhose English News Article articles datasets. The main reason is to balance the data, since OpenSources does not have a significant number of "reliable" news. The number of different domains included is 745 and the labels for each article correspond to its source's primary label in the OpenSources dataset.

2.8.3 Fake News Challenge Dataset

Although this dataset does not have annotations normally associated with fake news datasets, it was used for the Fake News Challenge Competition (mentioned in Section 2.7.1). The dataset has two different types of files. The first is composed of two fields, the id and corpus (body) of news. The second includes a news headline, a body id from the first file and a label regarding the stance of the headline (i.e. if it is related to the body). It is important to mention that each headline has multiple labels for different news corpus. The number of body ids for each headline is dynamic, ranging from 2 to 127. The main goal of this dataset was to create systems that were able to identify if a body agrees, disagrees, discusses or is unrelated regarding the headline. This methodology can be extrapolated to identify if a dubious piece of news is in accordance with what well establish news medium are publishing about the topic.

This dataset is an extension on the work of Ferreira et al. [20] and their Emergent Dataset which is described in the next subsection.

2.8.4 Emergent Dataset

The Emergent dataset [20] consists of rumors (in the format of claims) extracted by journalists from multiple sources. These claims are then manually linked to multiple news articles where each news article is labelled by journalists regarding their stance on the original claim. Finally, after multiple articles are collected, a veracity label is assigned. Three possible labels are possible for each news headlines: "for" (when the article is in accordance with the claim), "against" (when the article's statement is opposed to the claim) and "observing" (when no assessment of veracity is made by the article). The veracity label is initially set to "unverified". However, with the aggregation of articles

regarding a claim this label is converted to "true", "false" or remains "unverified" if no evidence is found.

2.8.5 Kaggle Dataset

This dataset [48] was originally posted in Kaggle[5], a platform for machine learning and data scientists enthusiasts, where datasets are published and machine learning competitions are held. The dataset uses OpenSources and each post is assigned their sources' label. The dataset contains approximately 13 000 posts from 244 different sources. Although this dataset follows the same methodology of the Fake News Corpus dataset, it adds a new label (bs) and a spam score provided by the bs-detector [62] application as well as some social feedback provided by the Webhose API[6].

2.8.6 Baly et al. Dataset

This dataset was used in the work of Baly et al. [6] and it was extracted according to the information available in MediaBiasFactCheck.com[7]. The labels provided on this website are manually annotated in two different categories. First, each news source is annotated according to its bias in a 7-point scale ("extreme-left","left","center-left", "center", "center-right", "right", "extreme-right"). Then, each source veracity is analyzed and categorized in one of three different labels ("low", "mixed", "high"). The dataset was built crawling the website and retrieving approximately 1050 annotated websites.

2.8.7 BuzzFeed Dataset

On August 8, 2017, BuzzFeed published a study referring to the partisan websites and Facebook pages in US politics and how they are proliferating after 2016 US presidential elections [55]. The dataset used for the analysis includes a collection of partisan Facebook pages and websites. These pages present elements of extremely biased opinions from the left and right political side. Although not all news stories included on the sources are false, they have an extremely biased opinion and

[5]http://www.kaggle.com
[6]https://webhose.io/
[7]https://mediabiasfactcheck.com/

were aggressively promoted and spread on Facebook. In addition to the Facebook page name and website, this dataset also includes information about the registered date of the websites and if these websites were linked to others (using Google Analytics and Google AdSense). In addition, engagement metrics regarding the difference Facebook pages (likes, reactions, shares and comments) is also presented. Each website is classified as "right" or "left" according to its political tendency.

2.8.8 BuzzFeed-Webis Fake News Corpus 2016

This dataset presents a set of 1627 news from 9 different publishers for 7 weekdays prior to the 2016 US presidential elections [46]. From the 9 publishers, 3 are mainstream (ABC, CNN, Politico) and 6 are hyperpartisan: 3 on the left (addicting info, occupy democrats, the other 98) and 3 on the right (eagle rising, freedom daily, right-wing news). Each news article was evaluated by journalists regarding their veracity in 4 classes: "mixture of true and false", "mostly true", "mostly false", and "no factual content".

2.8.9 PHEME Dataset

PHEME Dataset was first introduced by Zubiaga et al. [68]. The authors collected rumors from Twitter based on 9 newsworthy events classified by journalists. Then, after capturing a large set of tweets for each event, a threshold was defined and only tweets that had a substantial number of retweets were considered for annotation. The annotation process was conducted by journalists and each tweet was annotated as "proven to be false", "confirmed as true" or "unverified".

2.8.10 LIAR Dataset

The LIAR Dataset presented and described in [67] is, to the best of our knowledge, the largest human annotated dataset for false information analysis. The dataset is extracted from PolitiFact and includes 12.8K human annotated claims. Each claim is labelled with one of the six following veracity degrees: pants fire, false, barely-false, half-true, mostly-true and true. The dataset was also sampled to check for coherency. The agreement rate obtained using Cohen's Kappa was 0.82.

2.8.11 SemEval Dataset

One of the most recent datasets was created for the 2019 SemEval Task on Hyperpartisan News Detection [32]. This task consisted in detecting articles that followed a hyperpartisan argumentation. In other words, if it displays "blind, prejudiced, or unreasoning allegiance to one party, faction, cause, or person" [41]. This dataset is divided into two different types of annotations: by publisher (where the source of the news is annotated according to BuzzFeed journalists or the MediaBiasFactCheck website) and by article (annotated using multiple evaluations in a crowdsourcing fashion). The dataset has a total of 750 000 articles labelled according to their publisher and 645 articles labelled by Crowdsourcing workers. The labels by publisher are the same assigned in MediaBiasFactCheck.com. The labels assigned to the articles are just "true" (if the article is hyper-partisan independently of the political side) and "false".

2.8.12 NBC Twitter Propaganda Accounts Dataset

This dataset was collected by NBC news and refers to the 200k removed accounts that were publishing and spreading propaganda with the ultimate goal of influencing the 2016 US Presidential elections [44]. The list of users was released by the United States Congress thus it was possible to restore a part of the tweet data and users that were suspended by Twitter. The data is separated in users and tweets with information extracted from the Twitter API such as the number of followers, tweets, and favourites for the accounts and number of retweets, hashtags and mentions for the tweets. The main difference between this dataset and other is that it refers only to a type of tweets/users representing only one type of class/label.

2.8.13 Dataset Comparison

With a large diversity of datasets to tackle different tasks regarding the problem of detecting unreliable information, it is important to summarize the main characteristics of each dataset (like the content type, the last modification or what was their initial purpose). Table 1 presents an overall comparison among all datasets previously mentioned.

Table 1 Comparison between several state of the art datasets in the classification of false information

Name	Content Type	Last Modification	Number Entries	Human annotations?	Initial Task	Labels
OpenSources [40]	News Sources Websites	2017-04-28	833	Yes	Online Information Source	fake, satire, bias, conspiracy, rumor, state, junksci ,hate, clickbait, unreliable, political, reliable
Fake News Corpus [58]	News Articles	2018-02-13	9408908	No	Machine Learning/ Detection Systems	fake, satire, bias, conspiracy, rumor, state, junksci, hate, clickbait, unreliable, political, reliable
Fake News Challenge [18]	News Articles	2017-06-15	75390 (train+test)	Probably yes (it is not specified)	Stance Detection	agree, disagree, discusses, unrelated
Emergent [20]	News Articles	2016	262794	Yes	Stance Detection	stance labels: for, against, observing veracity labels: unverified, true, false
Kaggle Dataset [48]	News Headlines	2016-11-25	12999	No	Machine Learning/ Detection Systems	fake, satire, bias, conspiracy, state, junksci, hate, bs
Baly et al. [6]	News Sources Websites	2018-08-24	1066	Yes	Machine Learning/ Detection Systems	Factuality: Low, Mixed, High Bias: Extreme-Left, Left, Center-Left, Center, Center-Right, Right, Extreme-Right

(*Continued*)

Table 1 Continued

Name	Content Type	Last Modification	Number Entries	Human annotations?	Initial Task	Labels
BuzzFeed [55]	Facebook Pages	2017-03-31	677	Yes	Exploratory Analysis	left, right
BuzzFeed-Webis [46]	News Articles	2016-09-27	1627	Yes	Machine Learning/ Detection Systems	mixture of true and false, mostly true, mostly false, no factual content
PHEME [68]	Twitter Posts	2016-2-01	breaking news 9 conversational threads 330 tweets 4842	Yes	Exploratory Analysis	thread: rumor, non-rumor rumor: true, false, unverified tweets: support, denies, underspecified responses: agree, disagrees, comments
LIAR [67]	Claims	2017	12800	Yes	Machine Learning/ Detection Systems	pants-fire, false, barely false, half-true, mostly true, true
SemEval [32]	News Articles	2018/05/04	publisher 750 000 article 645	Partially	Machine Learning/ Detection Systems	article-hyperpartisan : true,false publisher-hyperpartisan: true,false publisher-bias: left, left-center,least,right-center,right
NBC Twitter Propaganda [44]	Twitter Accounts	2017-09-26	tweets 267336 accounts 513	No	News story	None

3 Discussion

Fake news is nothing new. It has been shown that even before the term has become trending, the concept has been active in different occasions. We might say that fake news is only a threatening problem these days because of the mass distribution an dissemination capabilities that current digital social networks have. Due to these problems, and particularly to the problems that consequently emerge from it for the society, the scientific community started tackling the problem, generally taking an approach of addressing first its different sub-problems.

The machine learning and deep learning approaches to the detection of fake content, or even to the correspondent network analysis to understand how this type of content can be identified, is diffuse and generally yet quite difficult to be understood by a general public.

Regarding the bot/spam detection, although we believe that they play an important role on the diffusion of fake news and misinformation, they do not represent all the accounts that spread this type of content. In some cases, the spreaders of misinformation are cyborg accounts (humanly operated but that also include some automatic procedures), as the authors in [11] refer. Another case which has been less discussed in the current literature are the human operated accounts that spread misinformation. Common examples are users who are highly influenced by extreme biased news and that spread that information intentionally to their followers. One could argue that this is the effect of the propagation of unreliable content through the network. However, the probability of having misinformation in our feed through the propagation of close nodes in our network is higher than from the original node that spread the content. Therefore, the evaluation of this accounts can be of major importance when implementing a solid system for an everyday use.

Another important aspect in adapting the current theory, to a system that informs users about which news are credible and which are misinformation, is the effect that such system may have on more skeptic or biased users. In fact, the "hostile media phenomenon" can affect the use of these systems if these are not equipped with the capability of justifying the credibility of the content. Hostile media phenomenon states that users who already have an opinion on a given subject can interpret the same content (regarding that subject) in different ways. The

concept was first studied in [1] with news regarding the Beirut massacre. Consequently, just like news media, such systems can be criticized by classifying a piece of news as fake by users who are in favor of the content for analysis. This leads us to the problem of the current detection approaches in the literature. For example, deep learning approaches and some machine learning algorithms are black-box systems that, given an input (in this case, a social media post), they output a score or a label (in this case a credibility score or a fake/true label). Therefore, explaining to a common user why the algorithm predicted such label/score can be a hard task. Furthermore, without some type of justification, such systems can be discredited. To tackle this problem in a real-world environment, the major focus after developing an accurate system must be to be capable to explain how it got to the result.

The new systems that detect, fight against, and prevent spreading of fake news/misleading news are every day increasing. Generically, we may say that all the stakeholders are delving into the subject, either from a theoretical or applyed, empirical, or even ad-hoc way. Most of these attempts use basic Artificial Intelligence from open source resources, in some cases paying for it (AIaaS). The automatic identification of Bots, fake accounts and deceiving information usually requires services like Entity Recognition, Sentiment Analysis, Relevance Identification, Importance Indexes, etc. In Table 2 we present some of the services available in the major providers of AI *as a service*.

These providers are nowadays in a privileged position to define 'facts' and reality according to their trained data sets and machine learning algorithms. This constitutes a dangerous situation because we are handing to these providers the power to define the reality.

4 Future Guidelines to Tackle Unreliable Content

We do believe that the analysis and effect of detection systems on the perception and beliefs of users towards fake news and all sorts of misinformation should be the next important step to be studied by the scientific community. Accordingly, we suggest some guidelines to approach the problem.

Table 2 Comparison between several AIaaS providers

Service	MS Azure [36]	IBM Watson [30]	Google Cloud [23]	AWS [3]	Text Razor [61]
Named Entity Detection	Yes	Yes	Yes	Yes	Yes
Sentiment Analysis	Yes	Yes	Yes	Yes	No
Event Detection	Yes	Yes	Yes	Yes	No
Relations between entities	No	Yes	No	No	Yes
Emotion Recognition	No	Yes	No	No	No
Importance Indexes	No	No	Yes	No	Yes
Image understanding	Yes	Yes	Yes	Yes	No
Key Phrases Detection	Yes	Yes	Yes	Yes	Yes
Language detection	No	No	Yes	Yes	Yes
Word Sentiment	No	Yes	Yes	No	No
Syntax Analysis	No	No	Yes	Yes	Yes
Does it have a free version?	Yes	Yes	Yes	No	Yes

In Section 2 we divided the current literature into eight different sections: unreliable account detection, content detection, network propagation analysis, fact-checking, content analysis, case studies, "fake news" detection systems and the datasets available to perform some of the tasks mentioned. In this section, we identify some of the limitations of current studies and future guidelines to tackle them.

Regarding unreliable accounts, the majority of works in literature have a focus on the detection of bots and spammers. However, not all unreliable accounts are restrained to these two groups. In addition, to the best of our knowledge, there are very few studies that analyze or try to detect accounts that publish or propagate unreliable content. One of the few examples we found was the work of Shu et al. [54] where they analyze the role of users profiles features for the classification of fake

news stories. Nevertheless, the task remains the same (i.e. classification of unreliable content). Thus, one of the suggestions for future research in this area is the analysis and classification of accounts according to the content that they spread. This problem can be modeled as a multi-classification task, where a social network account is classified with a single label. These labels can be similar to the ones used from OpenBias and MediaBias (for example "bias", "fake", "clickbait"). However, these labels are assigned to the accounts and not to posts/articles. Another way to define the problem is has a multi-label classification task where for each account, several labels can be assigned (based on the hypothesis that accounts can propagate several types of unreliable content).

With respect to content analysis, and to the best of our knowledge, only the work in [5] makes a clear comparison of the presence/absence of social feedback features when evaluating detection systems. In fact, if the major motivation from these studies is to develop a system capable of detecting unreliable content then, in a real world scenario, this system must be capable of detecting this information the earliest possible. Therefore, it is important to consider scenarios where no social feedback is present and how the different models would behave in those conditions. In addition, supervised systems in the unreliable content/"fake news" area should evaluate their performance based on metrics that considered how much information (and consequently time to retrieve that information) was used. In other words, it is necessary to obtain social feedback (or other types of features that take time to retrieve) to have a good prediction on a post, in a real scenario, this would "allow" the propagation of the unreliable content through the network. Thus, when evaluating the system, there should be some penalization for information and time that took to gather those features. We do believe that future research on this area should adapt time and information as two important metrics for assessing the quality/reliability of a system since these two variable are crucial in a real-world scenario.

Approaches presented in the fact-checking section are still at an early stage. However, some concerns arise from current studies. Namely, in stance detection, there are three main problems: the lack

of interpretability of the models [37], the slow training time and the possibility of the models being time-dependent. Once again, considering a real-world detection system, interpretability plays a key role for social network users with who suffer from confirmation bias (i.e. tendency to read, search and interpret information in a way that confirms its beliefs) to understand why certain content is unreliable. Time-dependent models and the slow training time are connected since the slow training time is not a major problem if the model is time independent. However, an analysis of the currently built deep models presented in Section 2.7.1 with recently extracted data is necessary to test this hypothesis. Therefore, an evaluation of the current models of stance detection with new and recent data should be performed to comprehend if the deep learning models developed are not overfitting to a particular dataset (like Emergent or the Fake News Challenge dataset). In the scenario where these models are time-dependent, slow training can be an issue in the implementation of a real-time unreliable detection system.

Table 1 shows that there is a large diversity of datasets available for the task of detecting unreliable content (and other similar tasks like stance detection). However, there are also some limitations that must be tackled. The first limitation is the generalization of unreliable websites provided by OpenSources and MBFC to unreliable articles. In other words, datasets like the Fake News Corpus, Kaggle Dataset and SemEval are annotated automatically based on the assumption that if a certain website is annotated with a specific label, then all articles of that website are labeled the same. This methodology has some limitations since, even websites labeled as false may provide true content. Therefore, human annotated articles provide more confidence in each entry than a generalization based on the source. However, there is a trade-off between human annotated entries and the size of the dataset since human annotation requires a large set of resources (for example for paying crowd-sourcing annotators or recruit a large set of volunteers) and large datasets are hard to label in a reasonable amount of time. A better solution seems to be tackling individual claims that were debunked by websites like Snopes and PolitiFact like the LIAR dataset. Thus, it would be interesting for future data extraction that

instead of only using the claims from those websites, to extract the original content from tweets and/or the original website to provide a multi-source and more enriched dataset.

5 Conclusion

Fake news is nowadays of major concern. With more and more users consuming news from their social networks, such as Facebook and Twitter, and with an ever-increasing frequency of content available, the ability to question the content instead of instinctively sharing or liking it is becoming rare. The search for instant gratification by posting/sharing content that will allow social feedback from peers has reached a status where the veracity of information is left to the background.

Industry and scientific communities are trying to fix the problem, either by taking measures directly into the platforms that are spreading this type of content (Facebook, Twitter, Google, for example), developing analysis and systems capable of detecting fake content using machine and deep learning approaches, or even by developing software to leverage social network users in distinguishing what is fake from what is real.

However, observing the various approaches taken so far, mainly by the scientific community but also some rumors about actions taken by Facebook and Google, we might say that mitigating or removing fake news comes with a cost [21]: there is the danger to having someone establishing the limits of reality, if the not the reality itself.

The trend to design and develop systems that are based on open source resources, frameworks or APIs which facilitate entity recognition, sentiment analysis, emotion recognition, bias recognition, relevance identification (to name just a few), and which may be freely available, or available at a small price, gives an escalating power to those service-providers. That power consists on their internal independent control to choose their machine learning algorithms, their pre-trained data and, ultimately, in a control over the intelligence that is built on the service provided by their systems.

Therefore, the saying "the key to one problem usually leads to another problem" is again true. However, we have not many choices at

the moment. Right now, the focus is to create systems that hamper or stop the proliferation of fake news and give back to the people, not only real information, but also a sentiment of trust in what they are reading. Meanwhile, we need to be prepared to the next challenge, which will be for the definition of what is important, or even more, what is real.

Acknowledgments

Nuno Guimaraes thanks the Fundação para a Ciência e Tecnologia (FCT), Portugal for the Ph.D. Grant (SFRH/BD/129708/2017). The work of L. Torgo was undertaken, in part, thanks to funding from the Canada Research Chairs program.

References

[1] Rasha A Abdulla, Bruce Garrison, Michael Salwen, Paul Driscoll, Denise Casey, Coral Gables, and Society Division. The credibility of newspapers, television news, and online news. 2002.

[2] Hunt Allcot and Matthew Gentzkow. SOCIAL MEDIA AND FAKE NEWS IN THE 2016 ELECTION. 2017.

[3] Amazon. Amazon comprehend. https://aws.amazon.com/comprehend/. Accessed: 2018-03-12.

[4] Qinglin Chen Mark Craft Anant Goel, Nabanita De. Fib – lets stop living a lie. https://devpost.com/software/fib, 2017. Accessed: 2018-06-18.

[5] Sotirios Antoniadis, Iouliana Litou, and Vana Kalogeraki. A Model for Identifying Misinformation in Online Social Networks. 9415:473–482, 2015.

[6] Ramy Baly, Georgi Karadzhov, Dimitar Alexandrov, James Glass, and Preslav Nakov. Predicting factuality of reporting and bias of news media sources. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP'18, Brussels, Belgium, 2018.

[7] F Benevenuto, G Magno, T Rodrigues, and V Almeida. Detecting spammers on twitter. *Collaboration, electronic messaging, anti-abuse and spam conference (CEAS)*, 6:12, 2010.

[8] Alexandre Bovet and Hernan A. Makse. Influence of fake news in Twitter during the 2016 US presidential election. pages 1–23, 2018.

[9] Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. Information Credibility on Twitter. 2011.

[10] Rogerio Chaves. Fake news detector. https://fakenewsdetector.org/en, 2018. Accessed: 2018-06-18.

[11] Zi Chu, Steven Gianvecchio, Haining Wang, and Sushil Jajodia. Detecting automation of Twitter accounts: Are you a human, bot, or cyborg? *IEEE Transactions on Dependable and Secure Computing*, 9(6): 811–824, 2012.

[12] Giovanni Luca Ciampaglia, Prashant Shiralkar, Luis M. Rocha, Johan Bollen, Filippo Menczer, and Alessandro Flammini. Computational fact checking from knowledge networks. *PLoS ONE*, 10(6):1–13, 2015.

[13] CNN. What we know about the boston bombing and its after math. https://edition.cnn.com/2013/04/18/us/boston-marathon-things-we-know, 2013. Accessed: 2018-06-12.

[14] Sarah Cohen, Chengkai Li, Jun Yang, and Cong Yu. Computational Journalism: a call to arms to database researchers. *Proceedings of the 5th Biennial Conference on Innovative Data Systems Research (CIDR 2011) Asilomar, California, USA.*, (January):148–151, 2011.

[15] David Conn. How the sun's 'truth' about hillsborough unravelled. https://www.theguardian.com/football/2016/apr/26/how-the-suns-truth-about-hillsborough-unravelled, 2016. Accessed: 2018-06-07.

[16] John P. Dickerson, Vadim Kagan, and V. S. Subrahmanian. Using sentiment to detect bots on Twitter: Are humans more opinionated than bots? *ASONAM 2014 – Proceedings of the 2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, (Asonam):620–627, 2014.

[17] Buket Erşahin, Özlem AktaÅŸ, Deniz Kilmç, and Ceyhun Akyol. Twitter fake account detection. *2nd International Conference on Computer Science and Engineering, UBMK 2017*, pages 388–392, 2017.

[18] FakeNewsChallenge.org. Stance detection dataset for fnc-1. https://github.com/FakeNewsChallenge/fnc-1, 2017. Accessed: 2018-04-12.

[19] Emilio Ferrara, Onur Varol, Clayton Davis, Filippo Menczer, and Alessandro Flammini. The rise of social bots. *Commun. ACM*, 59(7):96–104, June 2016.

[20] William Ferreira and Andreas Vlachos. Emergent: a novel data-set for stance classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1163–1168. Association for Computational Linguistics, 2016.

[21] Álvaro Figueira and Luciana Oliveira. The current state of fake news: challenges and opportunities. *Procedia Computer Science*, 121(December):817–825, 2017.

[22] Zafar Gilani, Ekaterina Kochmar, and Jon Crowcroft. Classification of Twitter Accounts into Automated Agents and Human Users. *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017 –*

[23] Google. Cloud natural language. https://cloud.google.com/natural-language/. Accessed: 2018-03-12.

[24] B Y Jeffrey Gottfried and Elisa Shearer. News Use Across Social Media Platforms 2017. *Pew Research Center*, Sept 2017(News Use Across Social Media Platforms 2017):17, 2017.

[25] Aditi Gupta. Twitter Explodes with Activity in Mumbai Blasts! A Lifeline or an Unmonitored Daemon in the Lurking? *Precog.Iiitd.Edu.in*, (September 2017):1–17, 2011.

[26] Aditi Gupta, Hemank Lamba, and Ponnurangam Kumaraguru. $1.00 per RT #BostonMarathon #PrayForBoston: Analyzing fake content on twitter. *eCrime Researchers Summit, eCrime*, 2013.

[27] Twitter Help. About verified accounts. https://help.twitter.com/en/managing-your-account/about-twitter-verified-accounts, 2018. Accessed: 2018-05-14.

[28] Alex Hern. Google acts against fake news on search engine. https://www.theguardian.com/technology/2017/apr/25/google-launches-major-offensive-against-fake-news, 2017. Accessed: 2018-04-13.

[29] Alex Hern. New facebook controls aim to regulate political ads and fight fake news. https://www.theguardian.com/technology/2018/apr/06/facebook-launches-controls-regulate-ads-publishers, 2018. Accessed: 2018-04-13.

[30] IBM. Ibm cloud docs natural language understanding. https://console.bluemix.net/docs/services/natural-language-understanding/getting-started. html. Accessed: 2018-03-12.

[31] Hamid Karimi, Courtland VanDam, Liyang Ye, and Jiliang Tang. End-to-end compromised account detection. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 314–321. IEEE, 2018.

[32] Johannes Kiesel, Maria Mestre, Rishabh Shukla, Emmanuel Vincent, David Corney, Payam Adineh, Benno Stein, and Martin Potthast. Data for PAN at SemEval 2019 Task 4: Hyperpartisan News Detection, November 2018.

[33] Bence Kollanyi, Philip N. Howard, and Samuel C. Woolley. Bots and Automation over Twitter during the First U.S. Election. *Data Memo*, (4):1–5, 2016.

[34] Sejeong Kwon, Meeyoung Cha, Kyomin Jung, Wei Chen, and Yajun Wang. Prominent features of rumor propagation in online social media. In *2013 IEEE 13th International Conference on Data Mining*, pages 1103–1108. IEEE, 2013.

[35] Iouliana Litou, Vana Kalogeraki, Ioannis Katakis, and Dimitrios Gunopulos. Real-time and cost-effective limitation of misinformation propagation. *Proceedings - IEEE International Conference on Mobile Data Management*, 2016-July:158–163, 2016.

[36] Microsoft. Text analytics api documentation. https://docs.microsoft.com/en-us/azure/cognitive-services/text-analytics/. Accessed: 2018-03-12.

[37] Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *CoRR*, abs/1706.07269, 2017.

[38] Damian Mrowca and Elias Wang. Stance Detection for Fake News Identification. pages 1–12, 2017.

[39] Richard Norton-Taylor. Zinoviev letter was dirty trick by mi6. https://www.theguardian.com/politics/1999/feb/04/uk.political news6, 1999. Accessed: 2018-06-07.

[40] OpenSources. Opensources - professionally curated lists of online sources, available free for public use. http://www.opensources.co/, 2018. Accessed: 2018-05-03.

[41] PAN. Hyperpartisan news detection. "https://pan.webis.de/seme val19/semeval19-web/". [Accessed: 2019-03-14].

[42] Verónica Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. Automatic Detection of Fake News. 2017.

[43] Stephen Pfohl, Oskar Triebe, and Ferdinand Legros. Stance Detection for the Fake News Challenge with Attention and Conditional Encoding. pages 1–14, 2016.

[44] Ben Popken. Twitter deleted 200,000 russian troll tweets. read them here., 2018. [Online; accessed 13-March-2019].

[45] Martin Potthast, Johannes Kiesel, Kevin Reinartz, Janek Bevendorff, and Benno Stein. A Stylometric Inquiry into Hyperpartisan and Fake News. 2017.

[46] Martin Potthast, Johannes Kiesel, Kevin Reinartz, Janek Bevendorff, and Benno Stein. Buzzfeed-webis fake news corpus 2016, February 2018.

[47] Jacob Ratkiewicz, Michael D Conover, Mark Meiss, Bruno Gonçalves, Alessandro Flammini, and Filippo Menczer Menczer. Detecting and tracking political abuse in social media. In *Fifth international AAAI conference on weblogs and social media*, 2011.

[48] Megan Risdal. Getting real about fake news. https://www.kaggle.com/mrisdal/fake-news, 2016. Accessed: 2019-03-14.

[49] Chengcheng Shao, Giovanni Luca Ciampaglia, Alessandro Flammini, and Filippo Menczer. Hoaxy: A Platform for Tracking Online Misinformation. pages 745–750, 2016.

[50] Chengcheng Shao, Giovanni Luca Ciampaglia, Onur Varol, Alessandro Flammini, and Filippo Menczer. The spread of fake news by social bots. *arXiv preprint arXiv:1707.07592*, pages 96–104, 2017.

[51] Chengcheng Shao, Giovanni Luca Ciampaglia, Onur Varol, Kaicheng Yang, Alessandro Flammini, and Filippo Menczer. The spread of low-credibility content by social bots. 2017.

[52] Prashant Shiralkar, Alessandro Flammini, Filippo Menczer, and Giovanni Luca Ciampaglia. Finding Streams in Knowledge Graphs to Support Fact Checking. 2017.

[53] John Merriman Sholar, Shahil Chopra, and Saachi Jain. Towards Automatic Identification of Fake News : Headline-Article Stance Detection with LSTM Attention Models. 1:1–15, 2017.

[54] Kai Shu, Xinyi Zhou, Suhang Wang, Reza Zafarani, and Huan Liu. The Role of User Profiles for Fake News Detection.

[55] Craig Silverman, Jane Lytvynenko, Lam Vo, and Jeremy Singer-Vine. Inside the partisan fight for your news feed, 2017. [Online; accessed 13-March-2019].

[56] Snopes. Fact-check: Comet ping pong pizzeria home to child abuse ring led by hillary clinton. https://www.snopes.com/fact-check/pizzagate-conspiracy/, 2016. Accessed: 2018-04-13.

[57] Kate Starbird, Jim Maddock, Mania Orand, Peg Achterman, and Robert M Mason. Rumors, False Flags, and Digital Vigilantes: Misinformation on Twitter after the 2013 Boston Marathon Bombing. *iConference 2014 Proceedings*, 2014.

[58] Maciej Szpakowski. Fake news corpus. https://github.com/several27/FakeNewsCorpus, 2018.

[59] Eugenio Tacchini, Gabriele Ballarin, Marco L. Della Vedova, Stefano Moret, and Luca de Alfaro. Some Like it Hoax: Automated Fake News Detection in Social Networks. pages 1–12, 2017.

[60] Marcella Tambuscio, Giancarlo Ruffo, Alessandro Flammini, and Filippo Menczer. Fact-checking Effect on Viral Hoaxes: A Model of Misinformation Spread in Social Networks. pages 977–982, 2015.

[61] TextRazor. Text razor – extract meaning from your text. https://www.textrazor.com/. Accessed: 2018-03-12.

[62] LLC. The Self Agency. B.s. detector – a browser extension that alerts users to unreliable news sources. http://bsdetector.tech/, 2016. Accessed: 2018-06-18.

[63] Robert Thomson, Naoya Ito, Hinako Suda, Fangyu Lin, Yafei Liu, Ryo Hayasaka, Ryuzo Isochi, and Zian Wang. Trusting Tweets : The Fukushima Disaster and Information Source Credibility on Twitter. *Iscram*, (April):1–10, 2012.

[64] Courtland VanDam and Pang-Ning Tan. Detecting hashtag hijacking from twitter. In *Proceedings of the 8th ACM Conference on Web Science*, pages 370–371. ACM, 2016.

[65] Chris J Vargo, Lei Guo, and Michelle A Amazeen. The agenda-setting power of fake news: A big data analysis of the online media landscape from 2014 to 2016. *New Media & Society*, page 146144481771208, 2017.

[66] Soroush Vosoughi, Deb Roy, and Sinan Aral. The spread of true and false news online. *Science*, 359(6380):1146–1151, 2018.

[67] William Yang Wang. "liar, liar pants on fire": A new benchmark dataset for fake news detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 422–426. Association for Computational Linguistics, 2017.

[68] Arkaitz Zubiaga, Geraldine Wong Sak Hoi, Maria Liakata, Rob Procter, and Peter Tolmie. Analysing how people orient to and spread rumours in social media by looking at conversational threads. In *PloS one*, 2016.

Biographies

Álvaro Figueira graduated in "Mathematics Applied to Computer Science" from Faculty of Sciences (UP) in 1995. He got his MSc in "Foundations of Advanced Information Technology" from Imperial

College, London, in 1997, and his PhD in Computer Science from UP, in 2004. Prof. Figueira is currently an Assistant Professor with tenure at Faculty of Sciences in University of Porto. His research interests are in the areas of web mining, community detection, web-based learning and social media automated analysis. He is a researcher in the CRACS/INESCTEC research unit where he has been leading international projects involving University of Texas at Austin, University of Porto, University of Coimbra and University of Aveiro, regarding the automatic detection of relevance in social networks.

Nuno Guimaraes is currently a PhD student in Computer Science at the Faculty of Sciences University of Porto and a researcher at the Center for Research in Advanced Computing Systems (CRACS – INESCTEC). His PhD is focused on the analysis and detection of unreliable information on social media. He had previously worked as a researcher in REMINDS project whose goal was to detect journalistically relevant information on Social Media. Nuno completed his master's and bachelor's degree in Computer Science at the Faculty of Sciences of the University of Porto. In his master's thesis, he developed a novel way to create time and domain dependent sentiment lexicons in an unsupervised fashion.

Luis Torgo is a Canada Research Chair (Tier 1) on Spatiotemporal Ocean Data Analytics and a Professor of Computer Science at the Faculty of Computer Science of the Dalhousie University, Canada. He also holds appointments as an Associate Professor of the Department of Computer Science of the Faculty of Sciences of the University of Porto, Portugal, and as an invited professor of the Stern Business School of the New York University where he has been teaching in recent years at the Master of Science in Business Analytics. Dr. Torgo has been doing research in the area of Data Mining and Machine Learning since 1990, and has published over 100 papers in several forums of these areas. Dr. Torgo is the author of the widely acclaimed Data Mining with R book published by CRC Press in 2010 with a strongly revised second edition that appeared in 2017. Dr. Torgo is also the CEO and one of the founding partners of KNOYDA a company devoted to training and consulting within data science.